Statistical Analysis Report for Final Analysis of NCIC CTG MA.27 Based on Statistical Analysis Plan V1.3 (October 26, 2010)

A RANDOMIZED PHASE III TRIAL OF EXEMESTANE VERSUS ANASTROZOLE IN POSTMENOPAUSAL WOMEN WITH RECEPTOR POSITIVE PRIMARY BREAST CANCER

Prepared by Judy-Anne W. Chapman, Ph.D., P.Stat. - November 8, 2010

This Final Analysis Report is prepared under NCIC CTG Standard Operating Procedure STG-SOP-0001 (V002).

PRIMARY OBJECTIVE:

The protocol-specified primary objective of this study is a superiority test comparison of event free survival (EFS) between postmenopausal women who have histologically or cytologically confirmed, receptor-positive, adequately excised, primary breast cancer and are treated with exemestane or anastrozole as adjuvant therapy.*

*Efficacy endpoint data, and at randomization stratification factor data are available for ITT assessment of the primary objective.

A section, **DATA QUALITY REPORT FOR THE FINAL ANALYSES** is included before the **STATISTICAL ANALYSES** section.

PRIMARY OUTCOME VARIABLE:

EFS, the primary endpoint of this study, is defined as the time from randomization to the time of documented locoregional or distant recurrence, new primary (contralateral) breast cancer, or death from any cause. For these analyses, all randomized patients are included, and any patients who did not develop the defined event at the time of the analysis, or were lost to follow-up, are censored at the time of their last contact. Patient data are analyzed in the groups to which patients were allocated, regardless of whether they received the assigned treatment.

SECONDARY OBJECTIVES AND OUTCOME VARIABLES:

Secondary endpoints include overall survival (OS), defined as the time from randomization to the time of death from any cause, and time to distant recurrence (DDFS), defined as the time from randomization to the time of documented distant recurrence. For the analyses of these time-to-an-event endpoints, all randomized patients were included and any patients who did not develop the defined event by the endpoint time of the final analysis or were lost to follow-up were censored at the time of their last contact. Patients were analyzed in the groups to which they were allocated regardless of whether they received the assigned treatment. The survival experience of patients in two treatment groups were described by the Kaplan-Meier method and a stratified log-rank test adjusting for the stratification factors at randomization was used to compare the two groups. An unadjusted analysis was also performed. As an exploratory analysis, a Cox proportional hazards model was used to adjust the observed treatment effect for the influence of various prognostic factors at study entry and identify factors significantly related to the survival outcomes.

For this analysis report, we provided by arm counts for contralateral breast cancer.

NEW EMERGENT OBJECTIVE AND OUTCOME VARIABLE:

An emergent role for operation of substantive competing risks (Chapman J, et al. J Natl Cancer Inst 2008; 100:252-260.) led to inclusion here of a new endpoint, disease-specific survival (DSS) defined as the time from randomization to death with or from breast cancer, including death following contralateral breast cancer. The analyses for this endpoint will be as above for other protocol-specified time-to-events.

TIMING OF THE INTERIM ANALYSES:

The first interim analysis was conducted to allow early termination of the study if the differences in efficacy, using a 2-sided test for superiority, were extreme between those assigned to Exemestane compared to those assigned to Anastrozole. This was performed in March, 2008 when a total of 277 events were observed. The DSMC recommended continuing the study after review of these results.

MA.27 data previously held by CTSU were brought to CTG for ongoing direct data management after the first interim analysis.

There was originally no provision in the protocol at either interim analysis for an examination of futility for the primary trial question of superiority, and one was not performed at the first interim analysis. However, a two-sided test for futility on the two-sided question of superiority was approved by CTEP and CIRB for the second interim analysis. **Note:** Neither the MA.27 trial statistician (JWC), nor the CTEP members of the NCIC CTG DSMC, were involved in setting the parameters for the futility analysis as they had seen the first interim analysis results.

The second interim analysis for event free survival and futility was to be performed on all randomized subjects when at least 430 events (2/3 of the total 644 events) were observed. Lan-DeMets error spending function was used to assess for superiority, and futility for superiority (Jennison C and Turnbull BW, Chapter 7). The early stopping boundary for superiority was based on a power family with power 3, which approximates the O'Brien-Flemming boundaries. The futility boundary was calculated using the stochastic curtailment method [Ware JH, Muller JE, Braunwald E. The futility index. Am J Med 1985;78:635–643. Lan KKG, Simon R, Halperin M. Stochastically curtailed tests in long-term clinical trials. Commun Statist C 1982;1:207–219.] with $\gamma = 10\%$. This approach was suggested by Freidlin and Korn (*Controlled Clinical Trials*, 2002) as a more conservative futility boundary. Those boundaries were used to reject the null and the alternative hypotheses, controlling the two-sided Type I error of 0.05 and the power of 80% at the end of the study.

If exactly 430 events were observed for the second interim analysis, the nominal critical points of the log rank test statistics for rejecting the null hypothesis and the alternative hypothesis would be 2.462 and 0.340, respectively. The corresponding two-sided p-values are 0.0138 and 0.734 . The null hypothesis would be rejected early due to evidence of superiority at the second interim analysis if the p-value of the stratified log rank test is less than or equal to 0.0138. The alternative hypothesis would be rejected for futility if the p-value was at least 0.734 , which corresponds to an observed hazard ratio of 0.968 for the second interim analysis. The nominal significance level for the final analysis would be 0.0448.

Should the futility boundary have been crossed, there were CTEP approved formal tests of pre-specified adverse events that were to be reported. Testing was with protocol-specified exact Fisher tests.

The clinical cut-off date for the second interim analysis was January 16th, 2009. All trial forms referring to data up to and including that date were collected by March 24th, 2009. Query resolution was continued under Biostatistical quality assurance review and resolution before an

April 2, 2009 final data lock for analysis. The DSMC recommended continuing the trial as planned.

DATA QUALITY REPORT FOR THE FINAL ANALYSES

The clinical cut-off date for the final analysis was April 15, 2010. All trial forms referring to data up to and including this date that were on forms dated up to the last data in house date of June 30th, 2010 are used for the final analysis. Query resolution was continued under Biostatistical quality assurance review and resolution before October 22, 2010 final data lock for analysis.

Data for the 749 IBCSG patients were included in the first and second interim analyses (11% of the previously planned 6840 protocol-specified North American patients), and are included in the final analysis.

Outstanding Queries (not returned): Total=1463

Baseline patient data (F1):309, of which 37 refer to adverse events.

Follow-up forms adverse event information (F5/F5S/F6): 447/58/2.

Event information at patient death/recurrence (F6/F9): 23/24.

(Overestimation on F6/F9 numbers because some of the query letters pertain to F6/F9 data that was dated after the clinical cut-off date (2010-APR-15).)

Queries returned, but not reviewed: Total=494

Baseline patient data (F1): 17, of which 2 refer to adverse events.

Follow-up forms adverse event information (F5/F5S/F6): 32/70/0.

Event information at patient death/recurrence (F6/F9): 0/0.

(Potential overestimation because query letters may have been processed, but date of completion not entered.)

Also, from 2003- 2008 (June), queries issued via CTSU RDC (OC) system cannot be tracked.

STATISTICAL ANALYSIS

Data considerations:

To compare EFS in the two arms of this study, a new variable was created which describes the length of EFS, and the SAS database was utilized to define a variable which describes whether the patient has had an event, or is censored. The analysis unit was days in order to minimize the number of ties in study follow-up time; 1 day was added to follow-up of patients who would otherwise have zero, for full ITT count at time zero.

Baseline characteristics of patients and accrual information for the study were summarized using SAS PROC FREQ. Patient characteristic data were provided on an ongoing basis by the central office, and their frequencies reported annually at the investigators' meeting, and semi-annually to the Data Safety Monitoring Committee (DSMC). The comparison of patient characteristics

between the two treatment arms was not the primary objective of the interim analyses so there were no formal general statistical testing for imbalances which may affect the primary endpoint; specific considerations of imbalances are included in the final analyses, in this report.

Toxicities were monitored on an ongoing basis by the central office, and their frequencies reported annually at the investigators' meeting, and semi-annually to the Data Safety Monitoring Committee (DSMC), the comparison of toxicities between the two treatment arms was not the primary objective of the interim analyses so there was no formal statistical testing for differences; such tests were performed in the final analysis. Reported adverse events were up to date to the clinical cut-off. Adverse events are classified for the final analysis to be for patients who received at least 1 dose of treatment, in the classifications by treatment received, and split into acute toxicity (within 30 days of treatment, or on-off treatment form) and delayed toxicity (> 30 days from treatment).

At the first interim analysis, the DSMC requested formal tests for differences in toxicity summaries; these tests were repeated at the second interim analysis, as well as at the final.

Additional CTEP approved formal tests of pre-specified emergent adverse events were performed at the second interim analysis, and repeated at the final analysis.

At the final analysis, new pre-specified tests were added to compare the trial arm experience of any clinical fracture, frailty fracture, and bisphosphonate use.

Toxicity tests utilized as treated populations in exact Fisher tests of differences by trial therapy.

Medical practice changes led to redesign of MA.27 during patient accrual/randomization:

1. MA.27 was originally designed as a randomized phase III trial of exemestane versus anastrozole with or without celecoxib in postmenopausal women with receptor positive primary breast cancer. On December 22, 2004 allocation of treatments was amended to remove celecoxib for safety reasons. This reduced the number of treatment arms from 4 to 2, i.e. exemestane + celecoxib; exemestane + placebo; anastrozole + celecoxib; anastrozole + placebo to exemestane; anastrozole. At that time, there was also a removal of the stratification factor for prophylactic aspirin use ($\leq$81 mg/day; yes, no), as this stratification factor was present only because of celecoxib. Celcoxib use was terminated immediately, while removal of stratification by aspirin use was implemented gradually through an approved protocol amendment.

2. The herceptin results presented at 2005 ASCO and implications for MA.27 patients led first to a memo June 6, 2005 that permitted herceptin use, and then to a protocol amendment. As of the November 24th, 2005 version of the protocol, text pertaining to herceptin use was included in protocol and in the data collection forms and ' herceptin use' became a stratification factor (yes,no).

The primary objective of this trial is to compare the efficacy of exemestane and anastrozole. The protocol indicated that a stratified two-sided log-rank test, adjusting for stratification factors, was the primary method to compare EFS between the two study groups. To handle the original randomization to celecoxib, celecoxib was used as a stratification factor.The stratification factors varied during the course of the trial and are summarized as follows:

1. Lymph Node Status: (positive/unknown/negative, for entire trial period)
2. Adjuvant chemotherapy: (yes/no for entire trial period)
3. Celocoxib use: (yes/no until December 22, 2004; unknown afterwards)
4. Aspirin use (yes/no, until after local centre change of protocol following December 22, 2004 discontinuation of Celecoxib; unknown thereafter)
5. Herceptin use: (unknown, before local centre implementation of November 24,

2005 protocol; yes/no afterwards).

The log-rank test has greatest power when the hazard ratio is constant (Cantor, 2003); we did not examine the assumption of proportional hazards in the interim analyses, where the trial data were likely under-powered for such an assessment. For the purposes of the interim analyses, we assumed no significant interaction effect for celecoxib in the assessment of exemestane and anastrozole (Green S, et al, 2002). Further, we neither tested nor adjusted for any imbalance in herceptin use when herceptin use was not a stratification factor. These items were considered in the final analyses.

A global unadjusted log-rank analysis was performed in the final analysis.

Further, the protocol stated that the survival experience of the patients in the two treatment groups would be described by the Kaplan-Meier method. There is no upper age restriction for entry of patients into MA.27; however, as the primary endpoint includes all types of death, it would not be necessary to consider the operation of competing risks for this endpoint (Chapman, et al, 2008). Should there be evidence of significant imbalances in factors, the imbalances may affect efficacy, so adjusted Cox survival plots would be generated.

The protocol indicated the performance of an exploratory Cox proportional hazards model analysis to adjust the observed treatment effect for the influence of various prognostic factors at study entry, and identify factors significantly related to EFS. There was no Cox modeling for the interim analyses. For the final analysis, the stratified Cox model was determined using a step-wise forward model building procedure, based on determination of a significant effect for a factor (two-sided $p \leq 0.05$) with likelihood ratio criterion test statistic ($\sim \chi^2_{(1)}$). The patient characteristics including baseline Raloxifene use were factors considered in those assessments. As described above, design changes since the trial's inception raise the potential of non-proportional hazards, a violation of the Cox model; these may accrue in particular due to the time periods when some patients received celecoxib or herceptin. We used cumulative hazard plots to graphically examine whether there was evidence of substantive non-proportional hazards specifically in the time-periods of celecoxib and herceptin therapy at the time of final analysis, which may affect the primary efficacy comparison of exemestane and anatrozole (Chapman, et al, 2008). Should there have been substantive evidence, particularly for trial therapy, then log-normal survival analyses would have been performed (Chapman, et al, 2008). We also looked for interactive effects between significant factors and trial therapy.

Other patient accrual timeline considerations which might have impacted efficacy:

1. MA.27 was activated in Canada in April 14, 2003, with ongoing accrual;
   in the US, June 30, 2003, with ongoing accrual;
   through IBCSG, starting September 30, 2005,
   with first patient accrued October 24, 2005 and
   accrual until December 15, 2006.

2. As of April 24, 2006, all North American patients had to be enrolled in a bone mineral density substudy, MA.27B, and a breast density substudy, MA.27D. IBCSG patients were not accrued to these companion studies.

3. MA.27B stratum A (BMD T-score $\geq$ -2.0 SD of mean value of peak bone mass in young normal women) was filled and closed September 22, 2006. Accrual to stratum B (BMD T-score < -2.0 SD of mean value of peak bone mass in young normal women) was open until May 30, 2008.

Tables and Figure included

Key results of the statistical analysis were summarized by tables listed below.
Most tables and the figures were generated directly from SAS output, and are attached with this plan.

Reporting of the results

The results of the interim analyses and other decisions regarding early termination of the study were referred to the NCIC-CTG Data Safety Monitoring Committee in a blinded fashion with separate sets of blinded codes for safety and efficacy endpoints, where there was previously blinding in the Spring Meeting Book Reports to the DSMC**.** Minutes of the meetings of the Data Safety Monitoring Committee were kept in our file. The decisions of the committee were communicated to the MA.27 Study Team. An outline of final analysis plans was submitted to the DSMC at the time of the interim analyses. The Executive Summary of the final analysis was sent to the DSMC.

REFERENCES:

Cantor AB. Extending SAS Survival Analysis Techniques for Medical Research, 2nd ed. Cary; SAS Institute, Inc.

Chapman JW, Meng D, Shepherd L, et al. Competing Risks of Death From a Randomized Trial of Extended Adjuvant Endocrine Therapy for Breast Cancer.
J Natl Cancer Inst, 2008; 100:252-260.

Freidlin B, Korn EL. A comment on futility monitoring. Controlled Clinical Trials 2002, 23:355-356.

Green S, Liu P-Y, O'Sullivan J. Factorial Design Considerations. J Clin Oncol 2002; 20:3424-3430.

Lan G, DeMets D. Discrete sequential boundaries for clinical trials. Biometrika 1983; 70:659-663.

Lan KKG, Simon R, Halperin M. Stochastically curtailed tests in long-term clinical trials. Commun Statist C 1982, 1:207–219.

Jennison C, Turnbull BW. Group Sequential Methods with Applications to Clinical Trials. Chapman & Hall, 2000, pg. 114.

Ware JH, Muller JE, Braunwald E. The futility index. Am J Med 1985;78:635–643.